Saumya Sinha | Curriculum Vitae

1300 30th Street, Apartment B4 -16, Boulder, CO 80303

+7204211144 • saumya.sinha@colorado.edu
https://www.linkedin.com/in/saumya-sinha-32799024/

I am a 2nd year CS Masters student at CU Boulder specializing in Machine Learning (ML). Prior to this, I have worked as a Data Scientist at American Express and Accenture for 3 years and have developed my expertise in ML, predictive modeling and statistical analysis. At CU, I work on Machine Learning, probabilistic models and deep learning. I am looking forward to research and make ML more adoptable in society and products.

Education

Academic Qualifications

- **University of Colorado, Boulder** — **Boulder, CO**
  Masters in Computer Science, GPA - 3.96/4 (3 semesters) — *2017-2019*
  Graduate Course Assistant for Machine Learning (Spring 2018)
  Teaching Assistant for Data Structures (Spring 2019, Fall 2018)
- **Indian Institute of Technology, Kharagpur** — **Kharagpur, India**
  Integrated BS+MS in Mathematics and Computing, GPA - 8.4/10, Department Rank - 5 — *2009–2014*

Projects

CU Boulder

- **Spatialized indicators of avalanche activity by SAR imaging and Machine Learning methods**

  Prof Claire Monteleoni, Sept 2018 - (Ongoing)
  Working to develop a Machine Learning based method for automatic detection of avalanche deposits in the Alps using the SAR data. Built modules to pre-process geospatial data, extract features in terms of slope, altitude, orientation and prepare an input image with labelled corridors corresponding to an avalanche event. Currently working with binary classifiers and Bayesian learning techniques so as to obtain a confidence over the results.

- **Augmented Reality based IOS app using text detection and manipulation for people with reading disabilities.**
  - Worked on Augmented Reality development using ARkit.
  - Represented virtual 3D text obtained from an deep-learning based OCR model and anchored it over the real text.
  - Added hit test (single/double tap)functionality on the virtual 3D text to navigate to a new screen or display objects relevant to the text. Also enabled anchoring the objects to a plane and interacting with it through hand gestures.

- **Analysis of comment removal on ChangeMyView Subreddit**

  Prof Chenhao Tan, May 2018 - July 2018
  Studied the removal of comments from /r/changemyview in terms of it's effect on both the participating users and the subreddit. Used Matching as a common causal inference technique, where a user whose comments have been removed forms a treatment group while a user whose comments have never been removed becomes a control user. Performed extensive comparison analysis of their commenting behavior and quality of posts across a certain period of time before and after the comment removal.

- **Building a GAN based creative assistant for Logo Generation**
  - Explored GAN generative models to create a creative assistance to help generate new and unique preliminary logo ideas.
  - Generated a set of semantically meaningful clusters that could improve in conditional GAN training.
  - Achieved best results (and mitigated mode collapse) with improved distance metric in the WGAN combined with ACGAN, conditioned on the above mentioned clusters.

- **Develop a scalable tool for Trend Classification in Twitter**
  - Built a Machine Learning pipeline with a Real-time Classifier which takes twitter data from a Kafka queue and writes into Cassandra. It was able to classify 6k tweets per minute with a pretty high accuracy(93%).
  - Tweets were classified into topics - Sports, Politics, Entertainment,Technology and Mood. Every trend was assigned a topic based on the majority classification of it's tweets.

- **Depth Map Estimation from Monocular Images**
  - Developed and compared various CNN based architectures from fully connected models to purely convolutional and also used transfer learning with pretrained models to estimate depth map from monocular images.
  - Used the depth map to identify and differentiate between the foreground and background of an image.
  - Visualised features formed by kernels.
  - Developed models on data with indoor scenes and objects. Achieved a very low RMSE on the object data, around 0.4.
- **Motion Planning with Rapidly Exploring Random Trees**
  - Explored and compared various variants of rapidly-exploring randomized-tree algorithm (RRT) to sub-sample the free space and find a path between two points in a 2D grid world with multiple obstacles.
  - Used GUI APIs (built by my team) to visualize and demonstrate the strengths and weaknesses of RRT, A-star and their variants like RRT*, Bi-directional RRT.
- **Natural Language Processing - Small Tools**
  - Implemented a Maximum Entropy Markov model for Name Entity Recognition task to find all the references to genes in biomedical journals.
  - Wrote a basic Word2Vec to learn word embeddings while building a sentiment classifier for hotel reviews.

IIT Kharagpur

- Detected and Mined communities in twitter, implemented clustering and graph partitioning algorithms after exploring social networking graphs and various interactions between users.
- Implemented Collaborative Filtering Algorithm for recommending brands given a user's brand choice from a fairly large dataset of user-brand transactions.

Previous Employment

- **American Express** — **Bangalore, India**
  Data Analyst, Big Data Capabilities and Algorithms division — *March 2016–May 2017*

  Text Mining/Modelling to build merchant database
  - Worked to improve data quality of merchant (businesses using Amex services) attributes like merchant names and addresses in a consumer-friendly database by integrating data from miscellaneous information sources.
  - Implemented Bayesian Probabilistic model (baseline) and predictive models using GBM to choose most accurate value for a given attribute and achieved 95% accuracy for most of the markets.
  - Tested out a sequence to sequence learning model using recurrent neural networks (LSTMs) for guessing the correct or cleansed values for the above textual attributes.

  Developed Training Module
  - Created training modules on Python, R, Hive and Unix, trained a group of (20-30) people on the above languages/tools.
- **Accenture** — **Bangalore, India**
  Data Analyst — *June 2014 - Feb 2016*

  Predictive Maintenance
  - Worked for an Oil and Natural Gas resources client, creating reproducible code pipelines for failure detection of equipments in the plant, including data preprocessing, exploration, feature extraction, modelling, diagnosis, performance evaluation and visualization.
  - Wrote R code for Predictive modeling using Machine Learning techniques – GLM, SVM, Random Forest and Unsupervised Machine Learning using Markovian modelling (for cluster analysis).
- **Environ Software (P) Ltd** — **Bangalore, India**
  Intern — *May 2013 - July 2013*

  Mathematical Modelling of the variation of tidal levels of a river
  - Devised and implemented Interpolation algorithms (Linear, Inverse distance weighted and Natural Neighbours interpolation) for triangular-mesh used in Finite Element method.
  - Formulated and Implemented (in C++) the numerical scheme to solve Saint-Venant Equations for the 1-D flow of the river, ran the simulation for 12 hours and successfully predicted the velocity, discharge and tidal levels of the river.

Technical skills

- **Programming Languages:** Proficient in: Python, C++, big data tools – Hive, PySpark, Hadoop MapReduce, Deep Learning Frameworks - Tensorflow, Keras, SKlearn
- **Statistical Skills:** R, SQL, SAS